DOCUMENT RESUME

ED 445 119 TM 031 803

AUTHOR Thompson, Tony D.; Davey, Tim
TITLE Applying Specific Information Item Selection to a Passage-Based Test.
INSTITUTION American Coll. Testing Program, Iowa City, IA.
PUB DATE 2000-04-00
NOTE 22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; *Reading Tests; *Selection; Simulation; *Test Construction; Test Items
IDENTIFIERS Dimensionality (Tests)

ABSTRACT

This paper applies specific information item selection using a method developed by T. Davey and M. Fan (2000) to a multiple-choice passage-based reading test that is being developed for computer administration. Data used to calibrate the multidimensional item parameters for the simulation study consisted of item responses from randomly equivalent groups of approximately 3,000 examinees each, each group taking 1 of 8 operational fixed forms from an existing paper-and-pencil test of reading comprehension. The main finding of the study was that the use of specific information item selection greatly aided efforts in creating a computer adapted test that met all the requirements of content specialists while at the same time controlling measurement precision. Although there were many complications in selecting the passage pool, it was worth the trouble because content specialists, under the current design, are able to review the combination of passages examinees can see beforehand, ensuring that each examinee will be administered a test that meets content requirements. Exposure control is done automatically at the passage level, and measurement precision can be specified because of the use of specific information item selection. The study provides evidence that the target information functions can be fairly well matched on average across the ability scale, but no evidence has been presented to show that this finding would hold for all simulated examinees within each score category. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 445 119

Applying Specific Information Item Selection to a Passage-Based Test

Tony D. Thompson
Tim Davey

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

T. D. Thompson

TM031803

This paper is prepared for the:
Annual Meeting of the National Council on Measurement in Education
New Orleans, LA - April 2000

BEST COPY AVAILABLE

Applying Specific Information Item Selection to a Passage-Based Test[1]

Tony D. Thompson
Tim Davey
ACT, Inc.

This paper applies specific information item selection to a multiple-choice passage-based test that is being developed for computer administration. The specific information item selection method is described in detail in Davey and Fan (2000), and so is only briefly described here. The method represents a practical alternative to standard maximum information item selection commonly used with adaptive tests. Although maximum information item selection maximizes the precision with which each examinee is measured, it does so at the expense of neglecting other important test characteristics. As described in Davey and Fan (2000), selecting items by maximum information has the potential disadvantages of variable measurement precision for examinees of the same ability, test measurement characteristics that are unduly dependent on the composition of the item pool, and a less balanced use of the item pool. The main feature of specific information item selection is that the highest discriminating items are reserved for those examinees that really need them, and consequently, target information functions can be matched with maximal use of the item pool. A further advantage of specific information item selection is that the measurement characteristics of tests are less dependent upon the composition of the item pool than with other item selection methods. This considerably simplifies the task of forming item pools, as it is not necessary to form strictly parallel pools.

Reading Test Content Specifications

Our focus is a test of reading comprehension we are currently designing as a passage-

---

[1] Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April, 2000. © 2000 ACT Inc. All rights reserved.

based, multiple-choice CAT. To describe how specific item information was applied in a computer-simulated version of the test it is first necessary to describe the content requirements of the test.

Content requirements specify that each examinee answer multiple choice questions associated with four reading passages. Each passage will have 15 items associated with it that will be pretested and we expect that at least 13 of these items will be judged acceptable for inclusion in an operational test. From this set of 13 items the CAT algorithm will select the items to be administered to the examinee for that passage. Passages are divided into four content types, and content constraints specify that each examinee 's test should contain passages from the four content types in a specified order. Thus, a passage from content type I is administered first, then a passage from content type II, etc. Although each of the four passages administered are from a different content type, the contents are related in such a way that the scores from passages 1 and 3 are combined into a subscore, and the scores from passages 2 and 4 are likewise used to form a subscore. The scores reported thus consist of an overall score and the two subscores.

In addition to having a passage from each of the four content areas, a number of formal and informal test construction rules need to be adhered to. For this reason, it would be preferable for content specialists to be able to review the passages received by the examinee ahead of time, to insure that all test construction rules are followed. However, using fixed forms would prevent adaptively selecting the passages. In choosing a CAT design, we considered three alternatives. One was to have the test administer preselected fixed forms, which is not a CAT at all of course. The second option would be to select passages and items in real time. In this case, each examinee would receive a set of passages that was best suited for their ability, and within each

passage, a set of items also tailored to their ability. A major disadvantage of this approach is that it prevents content specialists from reviewing forms ahead of time, and so would necessitate an extensive set of test construction rules to be encoded into the CAT passage selection algorithms. As noted before, the current fixed form test construction rules are complicated and in some cases not even fully formalized. Consequently, encapsulating these rules into computer code would be difficult, if not impossible. This leads to the third alternative for the CAT, which would be to fix the passage sequence but to have the items associated with each passage selected in real time. This was the option we chose, as it allows greater flexibility than using preselected fixed forms, and seems from the point of view of test construction to be more practical than selecting passages in real time.

Another design decision involved allowing examinees the opportunity to preview and review items within a passage. This was judged by content specialists to be essential given the nature of the test. However, by allowing item preview/review, the items of a passage must be administered as set, requiring the CAT algorithm to select all of the items corresponding to a passage prior to administering that passage. Another design decision was that we prefer to administer a fixed number of items to examinees due to speededness concerns.

Sample size limitations for pretesting will probably mandate that the algorithms that drive the CAT's item selection routines be based on a unidimensional IRT model. However, we have found that when conducting simulation studies it is more realistic to generate simulated data using a multidimensional model (Davey, Nering, & Thompson, 1997). Using a multidimensional model for the data allows one to test the robustness of the unidimensional model to realistic violations of model assumptions. The data used to calibrate the multidimensional item parameters for our simulation consisted of item responses from randomly

equivalent groups of approximately 3000 examinees each, each group taking one of eight operational fixed forms from an existing paper and pencil test of reading comprehension. A complete description of the data generation process can be found in the series of papers Nering, Thompson and Davey, (1997), Reckase, Thompson, and Nering (1997), and Thompson, Davey, and Nering (1997). The item pool for the simulation consisted of 32 reading passages and a total of 416 multiple-choice items.

Finding Acceptable Passage Sets

For purposes of controlling the frequency with which passages appear together, which may be thought of as a form of exposure control, we may choose to have anywhere from several dozen to more than one hundred passage combinations. Even with this many passages, content specialists will still be able to carefully review the suitability of each combination prior to the test administration. Our simulated pool has eight passages of each of the four content types so there are 8^4 (= 4096) possible passage combinations—many more than we need for purposes of passage exposure control. It makes sense then to choose passage combinations that are in some way optimal. One useful property we could require of our passage combinations would be for each set to contain a sufficient amount of information across the entire ability range, so that any examinee's ability could be well estimated regardless of the examinee's true ability. Passage sets that are low in information for certain parts of the ability continuum are not so desirable.

To operationalize the idea of meeting information constraints across the entire range of ability, the average information value for each of the two subscores was calculated across all of the 32 passages that formed our fixed form pool. These average information values represent our target information for the CAT. Ideally, every passage combination would contain enough information to match the average subscore information at every ability level.

The amount of unidimensional information that is obtained by a passage depends upon the particular items that are used with the passage. In our pool of 416 items, up to 13 items may be used with each passage. As stated previously, our simulated CAT will administer a fixed number of items to each examinee and these items will have to be selected by the CAT algorithm prior to administering the passage in question. Information values were obtained for each of the 8^4 passage combinations based on 8-11 items per passage. Seven ability values that spanned the range of ability were used to calculate the information, with the items being selected so as to maximize the possible information at each ability level. These values represent the amount of information that could be obtained if the ability parameter were known *a priori*, i.e., with perfect item selection. The number of passage combinations that meet or exceed the average information value for *all* ability levels is given in Table 1.

Table 1: Number of four-passage combinations that meet or exceed the average information values at all ability levels.

Number of items in passage	Number of combinations meeting restrictions
8	0
9	0
10	36
11	147

As can be seen from the table, 10 items per passage are required before any of the passage combinations meet the constraints for all ability levels, and even with 10 items per passage only 36 combinations exist.

It was thought that the number of required items could be reduced by adding a degree of passage adaptiveness to the CAT to allow the examinees' responses to influence the passage sequence administered. To implement this adaptiveness a multistage test was devised, wherein the first passage is administered to the examinee at random from the pool of passages eligible to

be administered in the first position. Then the examinee's ability parameter is estimated and compared to a cut score. The cut score determines which of two three-passage sets the examinee will receive to complete their test. We refer to the passage combinations in the multistage method a *passage set*. Each set has seven passages associated with it, including the routing passage and two three-passage sets, only one of which would be administered to the examinee. Using passage sets increase the number of possible passage combinations to 8^7 (= 2,097,152), and the number of these that meet or exceed the average information values at all ability levels is given in Table 2.

Table 2: Number of seven-passage combinations that meet or exceed the average information values at all ability levels.

Number of items in passage	Number of combinations meeting restrictions
8	59040
9	107442
10	199248
11	345548

Notice that with passage sets there are a large number of combinations meeting the information constraints, even with only eight items per passage. Yet, using passage sets will still allow content specialists to preview the tests prior to administration.

Although it not difficult to find all possible passage sets that meet the information constraints, a harder task is to form a pool of 40-50 passage sets so that every passage is equally likely to be selected. This is necessary to prevent the more informative passages from being overused. In a separate section given below, we present the details of an algorithm we developed to sort through the thousands of acceptable combinations and find a pool of passage sets that most nearly equalizes the frequency of passage administration. One finding that came out of the study was that no combination of forms was found that allowed all of the passages to be used

when each passage contained eight items. The best that could be done was for the first and third passages administered to contain nine items, and the other two administered passages eight items each. The number of acceptable passage sets found under these conditions was 104249.

CAT Administration Procedure

The previous sections detailed the steps leading up to the CAT administration procedure. We summarize these as follows.

1. Test developers set information targets that best complement the purposes of the test. In our case, the test purposes are consistent with the notion that all examinees at the same ability level should be measured to same degree of precision. This implies that maximum information item selection would not be appropriate for our needs.
2. A multistage testing model is selected to allow some degree of adaptive passage selection. The multistage testing model, along with test length, is selected so that test information targets can be met.
3. Passage set combinations are selected to form a pool. The selection is done so that each passage in the pool will be administered with approximately equal frequency.

With these preliminary steps completed, the actual process of administering the CAT may begin. The following seven steps outline the administration of the CAT reading test using the multistage method.

1. A passage set is selected at random from a pool of eligible combinations. The passage sets included in the pool would be carefully examined by content specialists to ensure that all content criteria were fully met. Also, the pool of passage sets would need to contain an equal representation of all of the available passages. This would build in a kind of exposure control for each passage.

2. A predetermined number of items from the first passage are selected at random. The items would be selected at random since the CAT algorithm would have no knowledge of the examinee's ability before the test begins. Another option would be to administer a predetermined set of items. Due to exposure control concerns, however, a better idea might be to use a number of predetermined item sets in such a way so as to ensure that all the items from the passage were used with equal frequency across examinees. The use of predetermined item sets would also reduce the possibility of a test being poorly matched to an examinee's ability, which would be more likely with random selection of items.

3. An ability estimate is determined from the examinee's responses to the items from the first passage and is compared to a cut score to select which of two three-passage sets is used to complete the test. We operationalized this in the simulation by numerically integrating the examinee's posterior ability estimate against the information functions of the two three-passage sets. The three-passage set with the greater potential information for the examinee's posterior is selected.

4. After the administration of each passage, the target information function for the subscore in question is updated. The update consists of reducing the target to account for the information already obtained. Then, an information target value for each subscore is obtained by numerically integrating each target information function over the posterior estimate of ability for the examinee. Thus, each subscore information target value is a scalar number that is essentially the expected target information for the examinee's ability estimate.

5. The item information function for each item associated with the second passage is numerically integrated over the posterior estimate of ability for the examinee. This gives an item

information value (a scale number) that represents the expected information of the item for the examinee in question.

6. A predetermined number of items, let us say *x*, are then selected to be administered. The set of *x* items selected is the one with item information values (step 5) that sum most closely to the subscore information target value (step 4) associated with that passage. In addition, some item level content constraints may have to be satisfied as well. This step required an integer programming problem to be solved.

7. Repeat steps 4-6 for the remaining two passages.

CAT Algorithms

Although the term CAT is often used rather generically, the performance of a CAT can vary greatly depending upon the particular algorithms used. The algorithms we used in the simulation follow those described in Thompson, Davey, and Nering (1998), in which a discrete item math test was simulated. For this simulation, the following options were employed using the 3PL model. The number of items administered was fixed, with the first and third passages containing nine items each and the other two passages containing eight each. The estimation algorithm for the provisional ability estimate was EAP, and the final ability estimate was computed by maximum likelihood. No exposure control was used at the passage level, as exposure control is enforced by the random assignment of passage sets to examinees. This assumes that the frequency of passage use is equally distributed throughout the passage set pool. Item exposure was controlled with the Sympson-Hetter (1985) method, except in the case of the first passage administered where the items were selected randomly. Although there exist several more sophisticated exposure control procedures, Sympson-Hetter is probably good enough for

our purposes since our primary concern is that the exposure rates of the reading passages are controlled for.

Finding the Ideal Passage Set Pool

A key step in implementing the routing passage CAT design was to insure that passages were administered with equal frequency. We specified the number of passage sets in our passage set pool to be 48. The number 48 was chosen because it was small enough that content specialists could still review all of the possible combinations of passages that could be administered to an examinee and would result in each of the 32 passages being used one-eighth of the time. The task was to select 48 passage sets out of the 104249 that met the information requirements so that when the passage sets are administered at random the distribution of passage use will be nearly uniform. Passage administration rates can be predicted by finding the marginal probabilities of administration for each of the two routing paths in a passage set. We refer to these probabilities as path probabilities. We estimated the path probabilities for each of the acceptable passage sets by administering the routing passage to 1000 simulated examinees whose ability parameters were drawn from a multivariate standard normal distribution. With the path probabilities in hand, it is easy to determine the administration rates for each passage for a group of passage sets.

We illustrate this computation by examining a single passage from a passage set. If the passage of interest is the routing passage, its administration rate for that passage set will be 1.0, as everyone receiving that passage set gets the routing passage. The administration rate will also be 1.0 in the atypical case that the passage appears in both paths. If the passage appears in a single path, the administration rate is simply the path probability. If the passage does not appear in the passage set its administration rate is 0. To calculate the overall administrate rate of a

passage for a group of passage sets, the administration rates for the passage are summed over the all of the passage sets in a group, and then the sum is divided by the number of passage sets in the group. In this manner the administration rates for each passage can be calculated for a given group of passage sets. For our test with 48 passage sets and where x_{ij} is the administration rate for the *i*th passage in the *j*th passage set, the average administration rate is given by,

$$\bar{x}_i = \frac{\sum_{j=1}^{48} x_{ij}}{48}.$$

The measure used to determine the degree of balanced achieved in passage administration rates was simply the sum of squared differences between the actual passage administration rates and the passage administration rates of a completely balanced pool. In the case of the reading test simulated, a perfectly balanced pool would use each passage one-eighth of the time. The equation stating the minimization criterion is,

$$\min \sum_{i=1}^{32} \left(\bar{x}_i - .125\right)^2,$$

where there are 32 passages available.

A least square difference rule seemed more appropriate than, say, a least absolute difference rule, in that a single large difference was a greater concern than several small differences. A large difference would indicate a passage being used either much more frequently or much less frequently than the other passages, raising exposure control concerns for that passage. Having small differences among the passage administration rates was judged to be acceptable.

Although the problem is well defined, finding the 48 passage sets that best balance the passage pool out of the 104,249 passage sets available is somewhat challenging. Examining all

possible combinations ($C(104249,48) \approx 10^{179}$) is not feasible. And standard optimization algorithms do not seem helpful due to the problem's non-linear nature. There seems no way to avoid the heavy combinatorics of problem as finding the optimal solution requires that passage sets be evaluated as a group rather than one at a time. This is analogous to the situation in stepwise regression, wherein adding one variable at a time to the model cannot guarantee that the optimal model will be discovered.

Although there is no procedure for finding the optimal solution, it is not necessary for us to find the absolute best solution. A solution that reasonably balances administration rates would be quite sufficient. To this end, we began a search for a heuristic algorithm that could find an acceptable solution in an efficient manner. One simple method would be to search the solution space randomly and take the best solution found. This method would work well if acceptable solutions occurred relatively frequently in the solution space. As we report below, however, even searching several billion random combinations failed to find an acceptable result. Instead, we began looking at algorithms somewhat akin to stepwise regression, something that combined the idea of a random search with the idea of evaluating passages sets for inclusion one at a time. The following outline briefly describes an algorithm that produces satisfactory results.

1. Select 48 passage sets at random from the pool available. The selected passage sets make up the *selected pool* and the remainder forms the *unselected pool*.
2. Proceed through the following steps.
    a. Cycle through all the passage sets of the unselected pool, starting in a random location. For each of the passage sets in the unselected pool, determine if swapping it with any of the passage sets in the selected pool would improve the

objective function. If it does then make the swap. If the swap does not change the objective function then make the swap with probability .5.

b. Take the current solution and compare it to the previous best solution. If the previous best solution is better, then use that as the selected pool.

c. For each passage set in the selected pool, replace it with a random passage from the unselected pool with probability .2. This step is only done every 15 iterations.

3. Iterate to step 2 as often as desired.

The algorithm above generally finds an acceptable solution after only a few iterations, but we let the algorithm run for several hundred iterations to try and find a better solution. Figure 1 gives the administration rates for the best solution that we have currently found to date. The administration rates vary from .090 to .149 as compared to ideal value of .125, and the total sum of squares value was .0096. Although we find this variation acceptable for our purposes, it is certainly possible that a superior solution exists. As a baseline of comparison, the best solution from a random search of several billion combinations had a sum of squares value of .1733 and administration rates that varied from .012 to .252. The algorithmic method not only found a vastly superior solution compared to the random search method, but it also took less computer time.

Matching Target Information Functions

A simulation study was conducted to examine the success of the specific item selection procedure to the passage-based test under examination. The study is not yet complete, and at this time only the match to target information will be presented.

Before discussing the results, though, we make a couple of notes concerning the figures. The results for the information functions described below are conditional on a unidimensional

approximation of true ability. The true ability approximation was constructed by first finding the unidimensional 3PL ability with response probabilities that best matched the response probabilities corresponding to the MIRT model that represented truth in the simulation (see Thompson et al., 1998). This was done for the overall score and both subscores using all of the items in the pool. The true ability approximations were then rescaled to true scale scores, using the same transformations that would be used for an operational test. The individual points in the plots represent 5000 simulees at each of the true scale score levels.

The match-to-target information functions are presented in Figure 2. The information plots in Figure 2 indicate how closely the information obtained in the CAT simulation matches the target information functions of the overall score and subscores. In addition to the target information functions, the plots also give the obtained 3PL information function of the CAT items using the best unidimensional ability based on the true MIRT ability. For the most part, the target information functions and the unidimensional approximation of the true information functions were quite similar, indicating that the CAT matched the targets on average.

Summary and Future Directions

Numerous complications surfaced during the design of the reading comprehension CAT. The main finding of the study was that the use of specific information item selection greatly aided the efforts of creating a CAT that met all the requirements of content specialists while at the same time controlling measurement precision. The requirements of the content specialists, particularly the need to review passage sets ahead of time and the desire to allow examinees to preview items within a passage, severely constrained the degree of adaptivity that could be implemented in the test. The lack of adaptivity in turn caused difficulties in selecting a passage set pool that insured that the target information function could be met for examinees of all ability

levels. And an algorithm needed to be devised to ensure that administration rates would be equally distributed among the passages.

We feel, however, that it was well worth the trouble to resolve these complications for the following reasons. Under the current design, content specialists are able to review the combination of passages examinees can see beforehand, which ensures that each examinee will be administered a test that meets content requirements. Examinees can preview/review items within a passage, an almost essential requirement for a test of reading comprehension where the stimulus is fixed. Exposure control is done automatically at the passage level. And measurement precision can be specified almost exactly due to the use of specific information item selection.

The previous point mentioned, that a target information function can be specified ahead of time and be matched precisely, has yet to be fully examined. This study gives evidence that the target information functions for the reading tests can be fairly well matched on average across the ability scale, but no evidence was presented to show that this finding would hold for all simulated examinees within each score category. We are actively investigating this question, as the success of specific information item selection depends upon the precise measurement of each individual examinee.

References

Davey, T., and Fan, M. (2000). *Specific Information Item Selection for Adaptive Testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA, April.

Davey, T., Nering, M.L., & Thompson, T. (1997). *Realistic Simulation Procedures for Item Response Data* (Research Report No. 97-4). Iowa City IA: ACT, Inc.

Nering, M., Thompson, T. D., & Davey, T. (1997). *Simulation of realistic ability parameters.* Paper presented at the Psychometric Society meeting, Gatlinburg, TN, June.

Reckase, M. D., Thompson, T. D., & Nering, M. (1997). *Identifying similar item content clusters on multiple test forms.* Paper presented at the Psychometric Society meeting, Gatlinburg, TN, June.

Sympson, J.B. & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thompson, T. D., Davey, T., & Nering, M. (1998). *Constructing adaptive tests to parallel conventional programs.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA, April.

Thompson, T. D., Nering, M., & Davey, T. (1997). *Multidimensional IRT scale linking without common items or common examinees.* Paper presented at the Psychometric Society meeting, Gatlinburg, TN, June.

Figure 1: Passage Administration Rates for Best Solution, Best Random Solution, and Ideal Solution

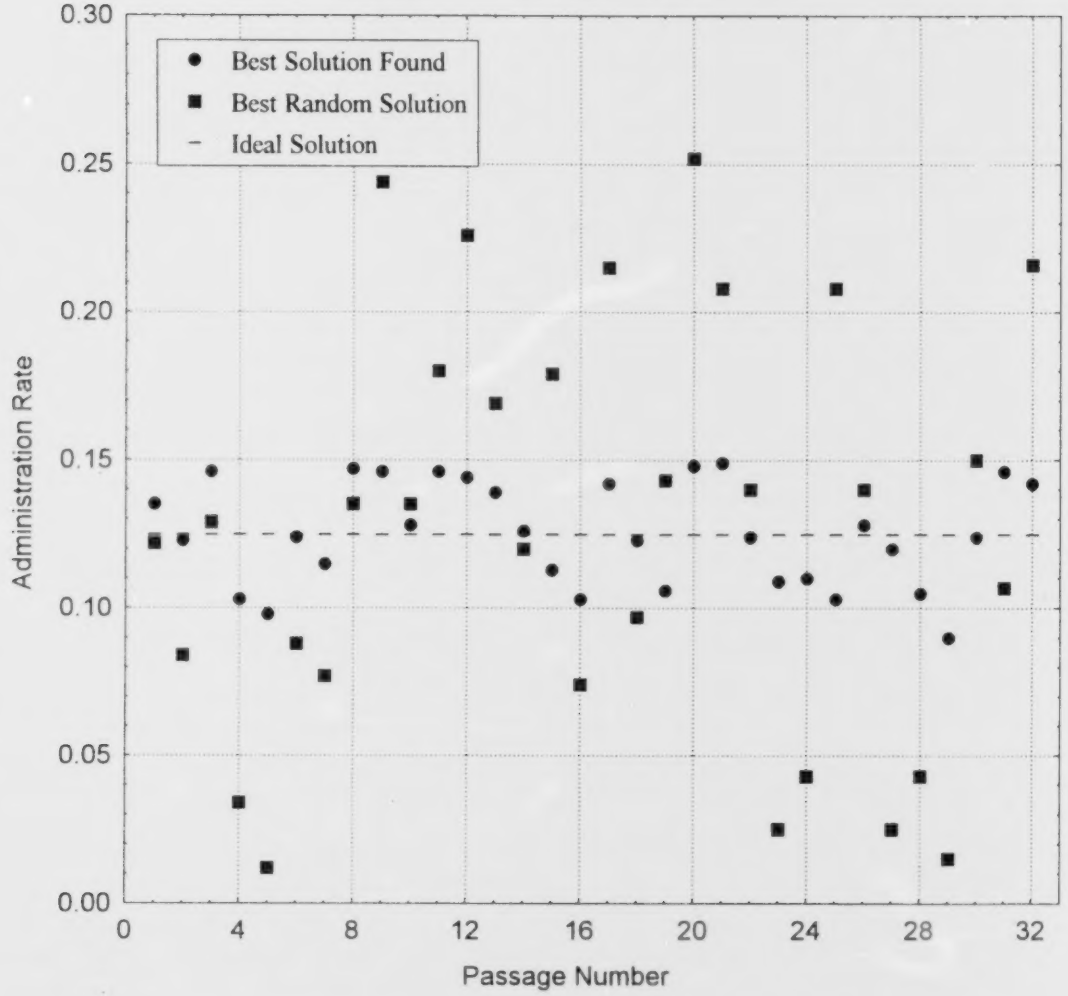
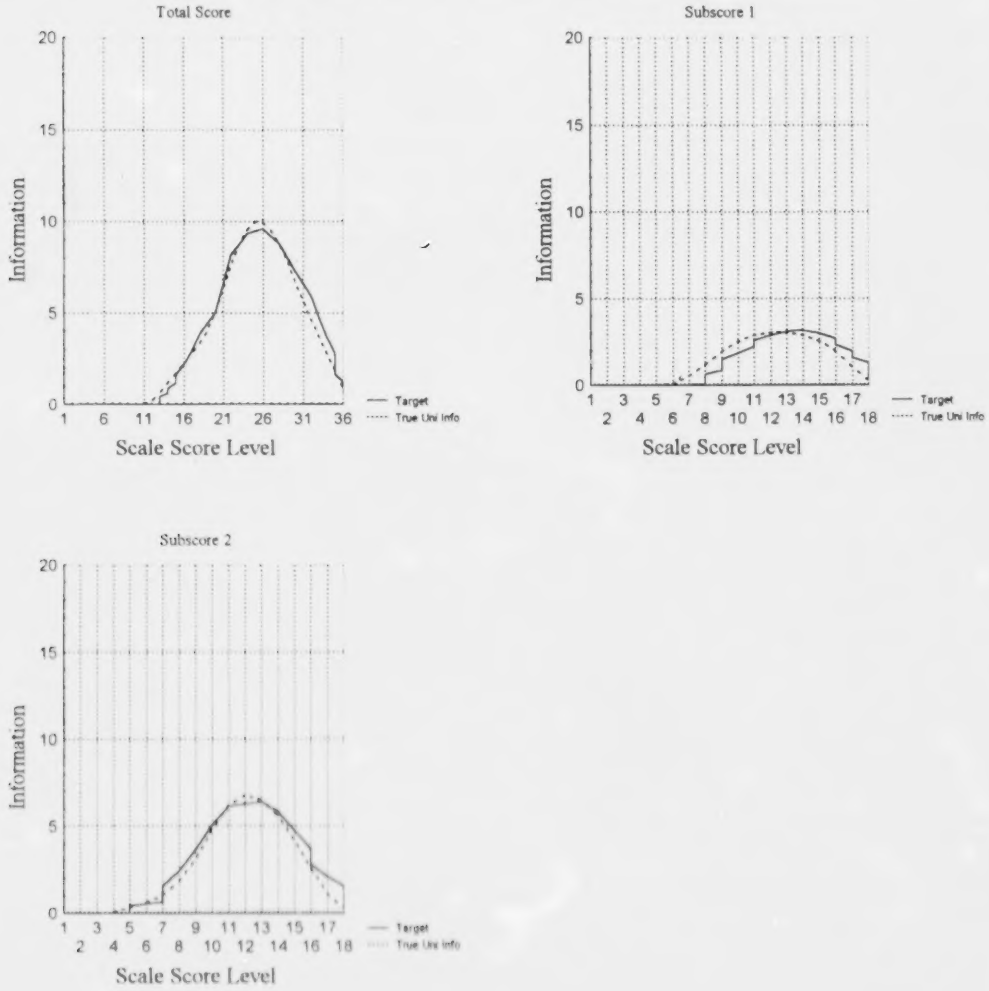


Figure 2: Target and Obtained Unidimensional Information Functions



BEST COPY AVAILABLE